

Users Guide on Scaled CMOS Reliability

Mark White
Mark Cooper
Allan Johnston

Jet Propulsion Laboratory
Pasadena, California

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

JPL Publication 11-12 11/11



Users Guide on Scaled CMOS Reliability

NASA Electronic Parts and Packaging (NEPP) Program
Office of Safety and Mission Assurance

Mark White
Mark Cooper
Allan Johnston

Jet Propulsion Laboratory
Pasadena, California

NASA WBS: 724297.40.49
JPL Project Number: 104593
Task Number: 40.49.01.07

Jet Propulsion Laboratory
4800 Oak Grove Drive
Pasadena, CA 91109

<http://nepp.nasa.gov>

This research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, and was sponsored by the National Aeronautics and Space Administration Electronic Parts and Packaging (NEPP) Program.

Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not constitute or imply its endorsement by the United States Government or the Jet Propulsion Laboratory, California Institute of Technology.

Copyright 2011. California Institute of Technology. Government sponsorship acknowledged.

TABLE OF CONTENTS

1.0	Introduction.....	1
1.1	General Characteristics of Scaling	1
1.2	Scaling Branches.....	5
1.3	Effects of Scaling on Reliability	6
2.0	Dominant Reliability Issues	7
2.1	Front-End Processing	7
2.2	“Back-End” Processing and Packaging	7
2.2.1	Metallization.....	7
2.2.2	Packaging	8
2.3	Methods to Evaluate Advanced Packages	9
2.4	Electrical Testing	10
2.5	End-User Reliability Testing	10
3.0	Commercial Practices.....	12
4.0	Screening Methods.....	13
4.1	Reliability Acceleration Factors	13
4.2	High-Temperature Testing (Burn-in).....	13
4.3	Electrical Testing	15
5.0	Derating Factors	16
6.0	Comparison of High-Reliability and COTS Manufacturers	17
7.0	Specific Recommendations	18
7.1	Relationship of Foundries and End-Manufacturers	18
7.2	Source Selection, Qualification, and Screening	18
7.2.1	Source Selection.....	18
7.2.2	Qualification Tests	19
7.3	Applications in Extreme Environments	21
8.0	Summary	22
9.0	References	23

1.0 INTRODUCTION

Reliability of advanced complementary metal-oxide semiconductor (CMOS) technology is a complex problem that is usually addressed from the standpoint of specific failure mechanisms rather than overall reliability of a finished microcircuit. A detailed treatment of CMOS reliability in scaled devices can be found in [1]; it should be consulted for a more thorough discussion. The present document provides a more concise treatment of the scaled CMOS reliability problem, emphasizing differences in the recommended approach for these advanced devices compared to that of less aggressively scaled devices. It includes specific recommendations that can be used by flight projects that use advanced CMOS. The primary emphasis is on conventional memories, microprocessors, and related devices. Field-programmable gate array (FPGA) and non-volatile memories are covered elsewhere [2, 3].

The discussion of reliability is limited to conventional CMOS devices, including the incorporation of strained silicon technology that allows bulk CMOS devices to perform better than silicon-on-insulator (SOI) technology. More advanced devices, such as dual gate CMOS and FinFet technology, are excluded.

Despite the aggressive advances in scaling in the commercial market, it is important to realize that the advanced devices used in space are more mature. It is highly unlikely that a space project would incur the high risk associated with the latest commercial technology, due to fundamental concerns about reliability. Furthermore, standard design practice for space systems requires de-rating of voltage, power, and temperature. Thus, many reliability issues that are important in commercial applications, and receive a great deal of attention in the literature, are of less concern in space.

1.1 General Characteristics of Scaling

Decreasing the feature size of CMOS devices not only allows more components to be placed on a single chip, but it increases performance by allowing faster switching (or clock) speeds, with reduced power compared to devices with larger feature size. Some general scaling trends for CMOS are shown in Figure 1.1-1; the values are taken from [4]. The horizontal scale is metal-oxide-semiconductor field-effect transistor (MOSFET) channel length, not feature size. For advanced devices, channel length is approximately 65% of the feature size used in processing.

Scaling theory originally had an objective of enhancing important circuit characteristics by decreasing effective MOSFET transistor channel length using constant electrical field strength as a guide or objective. In this manner, reliability would not be excessively compromised, speed would increase, and power dissipation per function would decrease. In order to accomplish this, however, power supply voltage (and other important voltages within the integrated circuit) would need to decrease by the same scaling factor. Some generations of scaling, however, do not decrease these voltages due to customer (and

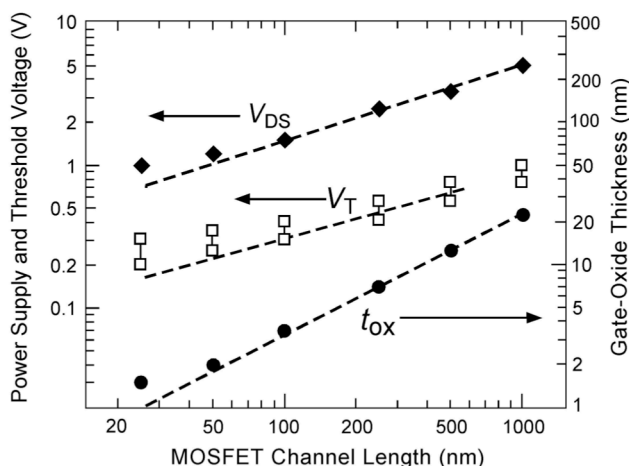


Figure 1.1-1. CMOS scaling trends for power supply voltage, gate threshold voltage, and channel length [4].

application circuit designers) insistence on compatibility with other integrated circuits in the application circuits, whose voltages were set by previous generations of large-scale integration/very-large-scale integration (LSI/VLSI) devices. Therefore, some technology nodes had compromised performance. Also, the higher electrical fields significantly compromised reliability since some failure mechanisms (such as time-dependent dielectric breakdown [TDDB]) are adversely affected (usually significantly and sometimes severely).

In addition, limiting power supply voltage (V_{DS}) decrease with smaller feature size, improves the inter-element spacing within individual transistors (the depletion width of a reverse-biased junction depends on voltage, and a lower voltage is required for reduced lateral spacing).

V_{DS} for advanced devices is seen by many VLSI designers to be limited to greater than 1 V, partly because of the need to maintain sufficiently large logic signals to provide noise margins compatible with other integrated circuits (which often operate at higher voltages), thus ensuring adequate design margin. Gate threshold voltage decreases with feature size, as shown in Figure 1.1-1. Although it is possible to reduce threshold voltage to about 0.25 V, higher values are required to be consistent with noise margin requirements as well as circuit requirements at higher temperatures. Most space systems require that circuits operate at junction temperatures of approximately 100–110°C, even though the actual junction temperature is usually lower.

The impact on circuit performance of CMOS scaling predicted by G. G. Shahidi in 2000 is illustrated in Figure 1.1-2, based on constant field scaling [5]. Scaling allows transistors to operate at higher speeds as the technology node (and channel length) advance with scaling. Function density is increased as more transistors may be fabricated in the same area (ref. circles in Figure 1.1-2) and power density also increases, but not as much. Note that the power supply is reduced as a result of scaling. The predictions made in 2000 agree reasonably well with the values implemented in devices with a feature size of 130 nm. Since that prediction, parts have advanced much further to features sizes as short as 22 nm. Constant field scaling has been modified, using different scale factors for the field in the gate and channel regions [6].

Power dissipation is a major concern for all integrated circuits. It often leads to higher junction temperature, which may result in sharply decreased reliability and lifetime. Package considerations, including thermal resistance, also play a role. The impact of scaling on power dissipation in advanced

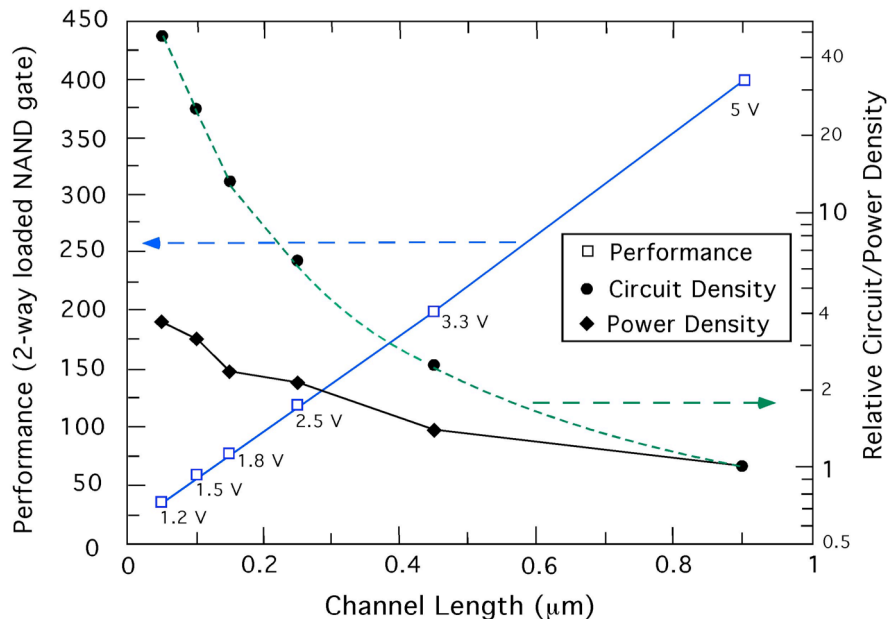


Figure 1.1-2. Scaling predictions in 2000 for advanced CMOS devices [5].

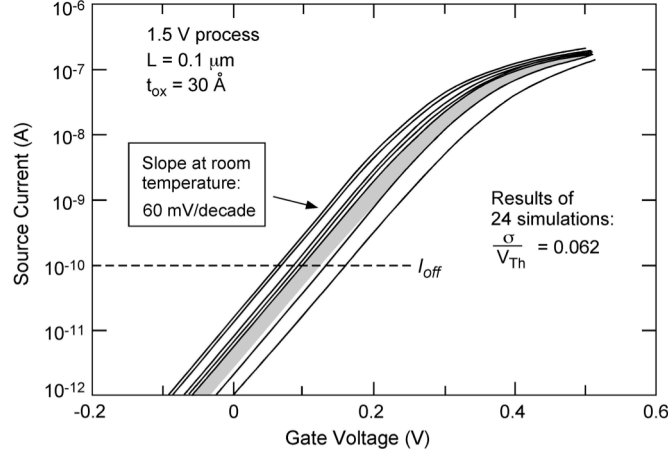


Figure 1.1-3. Sub-threshold slope showing the effect of statistical fluctuations in threshold voltage on the on/off current ratio. The slope is proportional to absolute temperature. A minimum threshold voltage swing of about 400 mV is required for an on/off ratio of 10^4 at high temperature. Additional noise margin requirements further increase this minimum value.

CMOS VLSI is shown in Figure 1.1-3 [7]. As transistor feature sizes are sharply reduced, dynamic power dissipation becomes a more important factor in total power dissipation. The two main aspects of power dissipation are the static or leakage power dissipation, dominated by sub-threshold current and tunneling through the gate region; and active power dissipation, dominated by capacitive load drive.

Static power dissipation becomes more important for scaled devices. The first factor is sub-threshold leakage current, nominally 60 to 75 mV/decade at room temperature. At 100°C , it increases to ~ 120 mV/decade for long-channel transistors. The difference between the “on” and “off” current is determined by the gate threshold voltage and the total logic swing. For high-performance devices, the ratio must be at least 10,000:1. Statistical variations in threshold voltage also affect this ratio when we consider the large number of devices on a chip; Figure 1.1-3 illustrates this for an advanced process [4]. Taking these factors into account, the total voltage swing for high-performance CMOS must be at least five times the sub-threshold slope value, or about 400 mV at high temperature. Other scaling branches (such as dynamic random-access memory [DRAM]) require much higher on/off current ratios, which can only be achieved with larger logic signals at the gate.

The above discussion applies to long-channel devices. As the channel length is aggressively shortened, the sub-threshold slope increases rapidly, limiting the minimum channel length. The other factor that affects static current is tunneling leakage through the gate oxide. For aggressively scaled high-performance CMOS, that current can be as much as 10% of the current during high-speed switching. It can be reduced by using thicker gate oxides, reducing the overall performance advantages of scaling, or by using different gate materials (such as hafnium dioxide) that have a higher dielectric constant than SiO_2 .

Active power dissipation, P_{act} , due to switching is approximated by the formula:

$$P_{\text{act}} = C_L f V_{\text{DD}}^2 \quad (1)$$

where the first term is the effective capacitive loading of each portion of the VLSI circuit, the second term is the switching frequency, and the last term is the square of the power supply voltage.

The user has little control over the capacitive loading since predominant capacitance is within the VLSI circuit. The user has less control over the power supply voltage internal to the VLSI circuit in that the majority of internal transistors are often powered by internally generated voltages, and these are only weakly dependent on the exterior power supply voltage. It is risky to overly reduce the exterior power

supply in the hopes of decreasing power dissipation since many internally necessary functions may not work correctly if that is done.

Reducing the exterior voltage to the lower end of the operating voltage range recommended by the part manufacturer is definitely not recommended for these reasons. Also, small decreases in voltage across the core transistors in the VLSI will have only a small effect on the active power dissipation.

The main impact that the user may have on the active power dissipation is to reduce the frequency of operation (f in Eq. 1). This has a linear effect on the active power dissipation.

A comparison of static and active power dissipation as feature size is reduced is shown in Figure 1.1-4. In this figure, leakage current at higher temperature is heavily influenced by drain-induced barrier lowering (DIBL), and becomes the dominant contributor below 50 nm. Dynamic (active) power increases more gradually with scaling.

However, there is an additional complication for threshold voltage scaling. As devices are scaled to small feature size, the number of dopant atoms in the channel decreases to the point that statistical fluctuations become important when we consider the distribution of threshold voltage over the large number of individual transistors on a large chip. The decrease in the number of dopant atoms for minimum geometry transistors is shown in Figure 1.1-5 [8]. This variance increases the variability of threshold voltage for highly scaled devices, which affects design margins and reliability. Manufacturers take this variability into account when they develop design rules for a specific process, but it has been a key stumbling block for highly scaled devices.

Gate oxide thickness has decreased more rapidly than the other parameters with scaling. Thin oxides result in higher transconductance (higher current drive), a key parameter in scaling metrics relating to switching. Surprisingly, the large decrease in gate oxide thickness has not directly impacted reliability. However, when the gate oxide thickness is reduced below 10 nm, a significant amount of leakage current can flow through the gate (due to quantum-mechanical tunneling). This does not directly affect reliability, but it increases the standby current. One approach to limit gate leakage is to use gate structures with higher dielectric constant (“high- k ” gates).

The reason that the reliability of scaled CMOS has remained acceptable is the sharp reduction of wafer defects in each technology node. This is illustrated in Figure 1.1-6 for 512 Mb DRAMs [1]. Similar improvements have been noted for advanced microprocessor technologies [4].

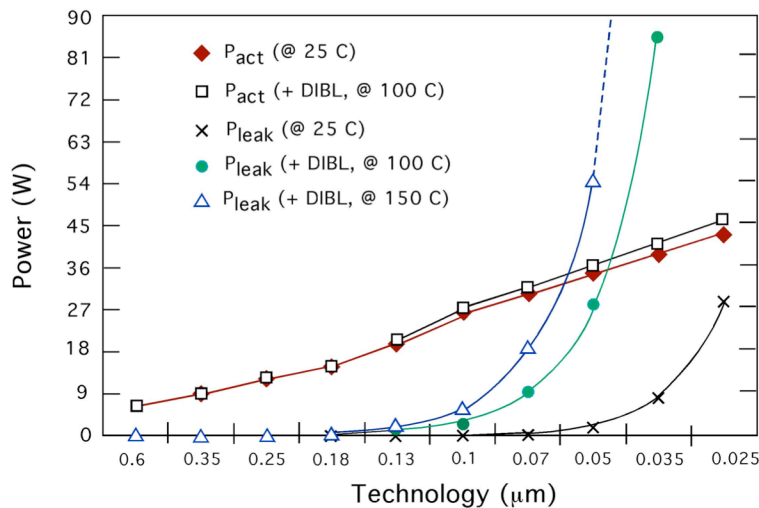


Figure 1.1-4. Scaling of active power and leakage current power with decreasing feature size [7].

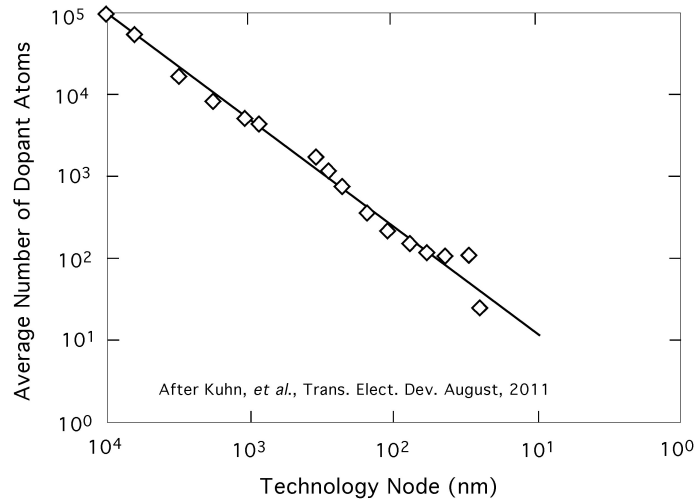


Figure 1.1-5. Decrease in the average number of dopant atoms for minimum geometry transistors as technology node scales [8].

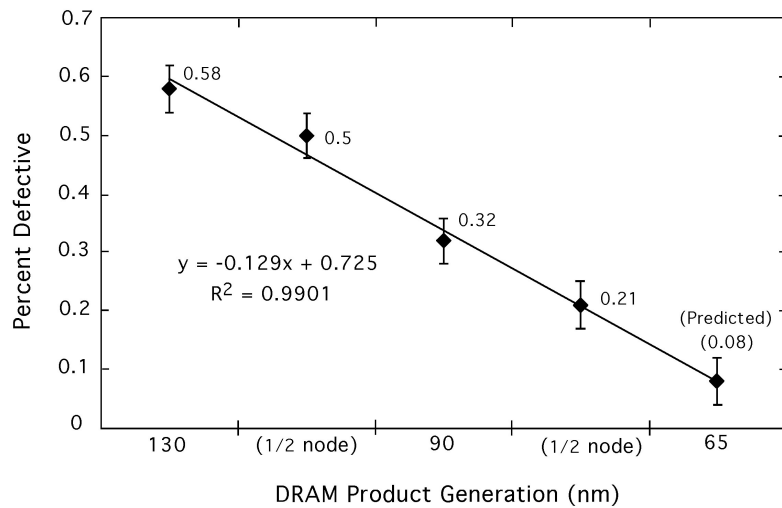


Figure 1.1-6. Decrease in defect density for 512 Mb DRAMs as the technology node is decreased [1].

1.2 Scaling Branches

During the last 15 years, a wider range of scaling rules has been developed, recognizing that different tradeoffs are needed for various end-products [4, 6]. For example, DRAMs need lower standby current and lower gate leakage compared to high-performance CMOS devices used in microprocessors and fast static RAMs. Consequently, thicker gates are needed for the internal pass transistors used in DRAMs compared to other CMOS processes. Table 1.2-1 shows general comparisons of the effects of scaling on these technologies.

On the other hand, special cooling methods can be used for high-end server applications, allowing more power dissipation compared to more normal applications of advanced CMOS. Scaling discussions often emphasize this branch, but it is clearly incompatible with most space applications. It is included in Table 1.2-1 for general comparison.

Table 1.2-1. Scaling trends for different CMOS applications.

Application	Primary Design Concerns	Power Supply Voltage	Gate Threshold Voltage	Reliability Issues
High performance	Switching speed	< 1 V	0.3	Electromigration Contact integrity
Mainstream	Maximum functionality density	1 to 1.2 V	0.4	Percent defect ratio of electrical test
Low power	Low total power dissipation	1.5 to 3.3 V	0.25	Percent defect ratio of electrical test
Memory	Low leakage and low standby current	3.3 V (internal back bias generator are often used to decrease leakage current)	0.5	Retention time for dynamic memories; early bit failures in large density SRAMs

Note further that modern CMOS technologies usually provide at least two different gate thicknesses, one for internal logic, which often operates at lower voltage with internally generated power, and one (or more) with thicker gate oxides for input/output (I/O) voltage that is compatible with 2.5 or 3.3 V external signals.

1.3 Effects of Scaling on Reliability

A recent publication from Intel [8] divides reliability issues associated with front-end processing into two categories: historical sources of variability, such as lithography, line roughness, and oxide thickness; and emerging sources of variability that are a direct result of scaling devices to very small dimensions. The main sources of variability are shown in Table 1.3-1. Most of the historical issues have been solved by clever improvements in processing. For example, the difficulty of coping with corner rounding and geometrical limitations associated with lithography has been overcome by designing devices in pairs for feature sizes of 65 nm and below.

Emerging issues are more difficult to deal with, and will be discussed in more detail in the next section.

Another key issue is that of device complexity. The large-scale devices that are the main focus of this document contain extremely large numbers of transistors. Special techniques (including on-chip error correction), and selective elimination of regions with defective devices that are identified during initial wafer probe tests, are used to improve manufacturing yield. Thus, there are aspects of the design that are quite different from older circuits.

Complex circuit design (e.g., modern SDRAMs, which are effectively small systems) must also be taken into account from the standpoint of reliability. Electrical testing is also a difficult problem that must take the design of specific circuit categories into account; it should also be tailored to reflect the actual conditions in National Aeronautics and Space Administration (NASA) applications.

Table 1.3-1. Effect of processing variations on reliability.

Historical Process Variations	Emerging Process Variations
<ul style="list-style-type: none"> • Patterning proximity effects • Line edge and line width roughness • Surface roughness (polishing) • Variations in gate oxide thickness • Fixed charge and oxide traps 	<ul style="list-style-type: none"> • Random dopant fluctuations • Variations in implants and anneals • Variations associated with strain • Gate material granularity

2.0 DOMINANT RELIABILITY ISSUES

2.1 Front-End Processing

The most important issues related to front-end processing are those involving the gate: random dopant fluctuations, discussed earlier, and the use of high-k dielectrics for applications that require the performance advantages of a thin gate oxide, but with lower gate leakage.

Significant process changes have become necessary to continue Moore's law scaling; examples include the use of hafnium oxide metal gate transistors (required by advances beyond the 65 nanometer technology node), and strained crystalline structure (required by advances beyond 90 nanometer technology node). The impact to the VLSI user is that new failure mechanisms and concerns are expected beyond the 32 nanometer technology node, and must be considered in manufacturer selection, flight lot qualification, and VLSI screening (see Section 7).

Time-dependent dielectric breakdown (TDDB), contact integrity, and hot-carrier degradation are typically less important. However, we should note that hot-carrier degradation has a negative activation energy, causing it to be more severe at low temperature. It may be of considerable importance for applications requiring extended operation at low temperature, such as surface exploration missions on the Moon or Mars.

Although there is considerable focus on mechanisms associated with front-end processing in the literature, there is little that the end user can do to deal with them.¹ Most manufacturers investigate these issues thoroughly, and ensure that their design rules and processing technology provide adequate reliability margins. Therefore, the main emphasis should be on other aspects of reliability, particularly those that are unique to NASA space applications.

2.2 "Back-End" Processing and Packaging

2.2.1 Metallization

There are many possible failure modes associated with metallization, particularly for processes that may use up to nine different metallization layers. Voids, grain boundaries, and thinning of metallization over non-planar regions are contributing factors, along with vias that are required to make connections between the different metallization levels. These mechanisms are somewhat intimidating, because a part can still function properly with localized defects or geometrical deficiencies. There is no obvious way to detect such defects in finished devices. Changes in their characteristics during extended operating periods can result in catastrophic failure.

Electro-migration is also an important mechanism, particularly for regions such as clock drivers and I/O circuits where higher currents are required. However, this is expected to be less important for devices used in space applications due to derating requirements that reduce the average current (note that dynamic current in CMOS scales directly with operating frequency).

Another issue is electro-migration from vias in copper interconnects. Low-k dielectrics are used in more advanced processes, and metal from the contacts can migrate within the dielectric materials. This process is quite different from CMOS processes with larger feature sizes, which do not use the new insulator materials [9]. Recent information on via reliability from a 32 nm process shows that this reliability problem changes in character for highly scaled devices. That study showed that only some of the vias actually failed, but that the ratio of those that failed—the number of fatal defects—increased

¹ Packaged devices do not provide direct access to internal transistors, limiting the ability to examine most of the mechanisms associated with front-end processing. For example, hot-carrier degradation is usually investigated with test transistors, applying much higher voltages to the drain and gate. Most packaged devices incorporate overvoltage protection that limits the maximum voltage. There are other cases, such as DRAMs, where the internal voltage used for access transistors is derived internally; it is unaffected by the external power supply voltage.

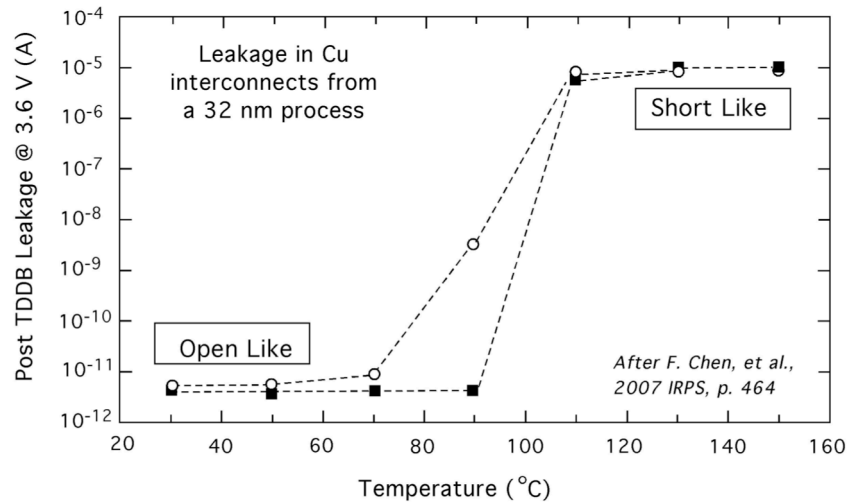


Figure 2.2-1. Metal-insulator transition in the low-k dielectric material used for interconnects in an advanced 32 nm process [9]. The leakage takes place between the materials used in multi-level metallization regions.

relative to the total number of vias as the area of the via was reduced. The study also showed that if the power through the vias was too high, a transition from insulator to conductor could take place, even at much lower temperatures than the typical temperature required for such transitions. This is shown in Figure 2.2-1. It illustrates that new reliability problems can be expected as devices are scaled below the 45 nm node.

2.2.2 Packaging

A number of package types are used in advanced VLSI/CMOS. Conservatism of the space community notwithstanding, advanced VLSI devices may not be available in traditional military packages. These military packages have a large database and heritage in space applications, and therefore, have lower risk in future space missions. VLSI parts in dual-side flatpacks and dual-inline packages (DIPs) are now becoming rare. Four sided flatpacks (quad flatpacks) using hermetic (ceramic) technology are also becoming rare. Commercial technology is driving toward a more I/O efficient usage of pinout. Area arrays (ball grid arrays, pin grid arrays and column arrays) are becoming the industry standard. Since the commercial marketplace is driven by high volume and low cost, advanced technology devices may only be available in non-hermetic plastic packages, which may be built in large batches at lower cost and with more uniformity.

Reliability may be adversely affected since the plastic package is hygroscopic and therefore will allow moisture to transport through the plastic into the die region (including the bond wires). For this reason, the plastic must be optimized to minimize transport of chemically active (and deleterious) ingredients such as chlorides and ionic contaminants (sodium and potassium).

There is also more stress on the die within the package due to mismatch in the thermal expansion coefficient of the die and package. That not only limits maximum and minimum temperature, but may also affect screening methods such as the temperature cycling traditionally used in burn-in and qualification.

Ball grid arrays are used for large-scale devices with high pin count. In such packages, the semiconductor chip is “flipped” and placed over a carrier with an array of solder balls. The balls are aligned with contact regions on the inverted chip. Hermetic packages are not available. More stress occurs at the corner regions of these packages compared to the center, which can cause cracking as well as intermittent or open contacts.

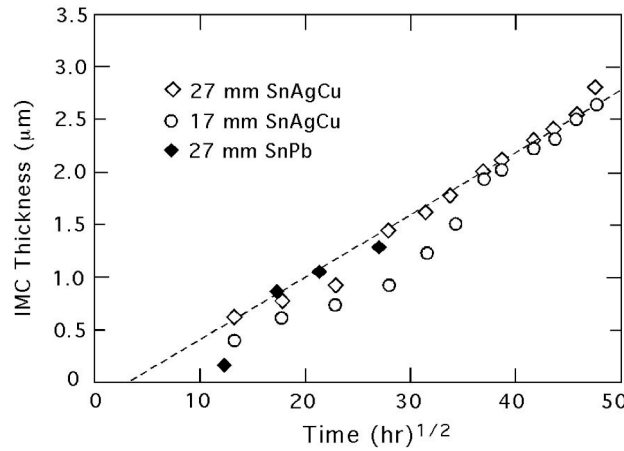


Figure 2.2-2. Increase in the buildup of intermetallic growth compounds in column grid arrays [10]. The results fit a diffusion model, and are more directly applicable to failure modes expected in conventional space applications where deep thermal cycling is not expected.

Column grids, which use small columns of material to establish contact between the package and device contacts, have also been developed. Column grid arrays provide a broader contact area compared to ball grid arrays, and are preferred for space products. However, they are only available for some devices.

Ball and column grid arrays provide an interesting conflict regarding testing and screening. One method to evaluate their reliability is to subject them to a series of thermal cycles, using the number of such cycles before failure as a metric. Although this is probably a good way to evaluate these devices for a Mars surface application where daily thermal cycles occur, it may be ineffective for more conventional space applications where only small thermal cycles take place.

A different approach is to evaluate the thickness of intermetallic growth in contacts (at a constant temperature), which is a more likely failure mechanism for conventional space applications. Recent results for three different ball grid designs are shown in Figure 2.2-2 [10]. The results fit a diffusion model, and provide a better way to evaluate this particular failure mechanism compared to deep thermal cycling.

2.3 Methods to Evaluate Advanced Packages

Failure rates in packages are found to scale as the range of temperature cycling according to the Coffin-Manson equation [11]:

$$FR \sim (T_u - T_l)^q \quad (2)$$

The temperatures are the upper and lower temperature of the temperature cycling testing, and the exponent q is a material-dependent parameter that must be established empirically. Typically, q is greater for more brittle materials such as plastics (a number such as 6 is typical) and is lower for a more ductile material such as solder (a number such as 2 is typical).

For advanced packages, the upper and lower temperature may be restricted (for military packages, the upper and lower temperature have been -65°C and $+150^\circ\text{C}$ respectively). Most recent generations of packaging have the acceptable lower temperature of -55°C and $+125^\circ\text{C}$. Advanced technology packages may have even more restricted temperature ranges.

Industry data is available to establish the important exponent q . In addition, correct values of this parameter should be sought from the part manufacturer.

In addition to temperature cycling experiments (either by the part manufacturer or user), a large sample Destructive Physical Analysis (DPA) is advisable to ensure that all failure mechanisms associated with advanced technology packages are found, and the relative failure rate and number of cycles to failure is established for new generation packages.

These issues become more critical when the package has not been used in space applications of similar duration and environments to those anticipated in the applicable space mission. Temperature cycling data to significant levels of failure (such as 10–25% of the packages), or for at least 1000 and preferably 2000 temperature cycles, should be captured from published literature or part manufacturer testing. Data taken at different temperature ranges may be reasonably interpolated using the Coffin-Manson equation.

Failure analysis should be done on all failures of these extended temperature cycling experiments. This will determine if new failure mechanisms have appeared in this generation of packaging. Appropriate screening methods to eliminate early temperature cycling failures must be found to determine the ability of the advanced packages to withstand the anticipated mission thermal environments.

2.4 Electrical Testing

Although it was not discussed earlier, electrical testing is one of the most important reliability issues. Extremely large numbers of individual transistors can be placed on a single die, producing very complex devices. It is extremely difficult to develop test methods for such complex parts that are capable of verifying that the device functions properly in all operating modes.

A great deal of effort has been expended to develop test patterns that can verify overall functionality, as well as sensitivity to different logic patterns and timing. Manufacturers often develop special test modes that can be used more efficiently. It is often possible to get such information, and incorporate these tests into the overall test approach for packaged devices.

When such data are not available, and the device is topologically complex (not true of most memories, true of microprocessors), then serious consideration should be given to testing the VLSI parts in an application environment or similar.

2.5 End-User Reliability Testing

Fundamental reliability mechanisms have to be solved at the manufacturing level. The end user can perform additional tests on assembled parts to further improve reliability, and perhaps identify marginal devices. However, these tests are quite limited, and tend to be focused on mechanisms associated with packaging. Examples include:

1. Electrical testing, which may include specific conditions that cover the expected applications
2. X-ray examination of devices to examine die attach, bonding, and other factors that might affect the ability of a part to withstand shock and vibration, or to dissipate heat
3. Burn-in tests to eliminate devices that are subject to infant mortality. Although burn-in is often used for devices with hermetic packages, the maximum temperature that can be used for non-hermetic packages is usually too low to make such testing effective. Therefore, consideration should be given to using voltage acceleration factors. However, burn-in at higher than normal voltages should still be within the part manufacturers recommended operating voltage range to prevent damage to the device.

In advanced technology packages, the glass transition temperature may be less than traditional values (typically +150°C). The glass transition temperature is the value (or range) where the temperature coefficient of the epoxy mold compound increase by a significant factor (often 3–4) making the mismatch within package elements much more severe. This means that repetitive temperature cycling close to or through the glass transition temperature is damaging to plastic packaged parts or parts with epoxy enclosed within them. Typically, cracks and separations are induced by this stress leading to early time

reliability issues. One time stress at a steady state elevated temperature such as during burn-in or life test is not deleterious to plastic packaged parts, as demonstrated by copious life test data generated at the Jet Propulsion Laboratory (JPL) and other centers. However, for these reasons mission operation near the glass transition temperature is discouraged.

3.0 COMMERCIAL PRACTICES

The commercial market drives advanced technology VLSI. Wafer fabrication is very expensive, with modern wafer fabs costing upwards of several billion dollars. Therefore, commercial demands control wafer fabrication processes. Design rules are set in the wafer fab by evaluation of advanced technology devices using test structures built into each wafer. Test structures are specifically designed to emphasize one failure mode of concern. For example, electro-migration test structures use serpentine metallization patterns at the extremes of the allowable geometries and surface complexity (e.g., oxide steps).

Commercial manufacturers assume that if test structures data, using commonly determined acceleration factors, show reliability greater than the market-driven needs, then the device may proceed to prototyping. Rarely do such prototypes fail reliability testing such as burn-in and life test. Commercial manufacturers assert that infant mortality is no longer a concern and has been removed by proper application of design rules. In those devices where some infant mortality is still anticipated (such as dense dynamic and static RAMs), experiments are conducted to establish the minimum burn-in time necessary. In fact, typically the burn-in time is aggressively shortened as the technology matures based on frequent life test experiments.

Automatic electrical testing is a major component of the end cost of manufacturing VLSI devices. In some cases, 30% of the cost of producing such devices is subsumed within the design and implementation of 100% testing. VLSI testers routinely cost several million dollars, and many are required to implement production test. Therefore, commercial manufacturers aggressively reduce test time and often use single temperature testing with guard bands to ensure (to their satisfaction) that three temperature electrical characteristics are met.

4.0 SCREENING METHODS

Special test and screening methods are typically required before semiconductor devices can be used in space. When parts are procured from a space-qualified manufacturer this is usually done at the factory, although the tests that are done as part of the normal process do not necessarily comply with NASA requirements. Thus, additional tests may be needed in order to qualify specific lots of devices.

For parts that are not space qualified, test and screening methods must be done on devices after final packaging. The specific requirements are included in NASA and JPL part requirements [12, 13].

Although screening methods are reasonably well established for conventional parts, they must be modified to some extent for devices with very small feature sizes. From the discussions in Sections 1 and 2, the best strategy is to rely on manufacturers to deal with front-end processing mechanisms, as well as those associated with metallization, interconnects, and bonding. The net recommendation is to concentrate on screening methods that are related to packaging, electrical performance, and unique requirements for specific applications, such as extreme temperature, or extended thermal cycling.

4.1 Reliability Acceleration Factors

Screening and testing methods for reliability mechanisms usually assume that the failure mechanism can be accelerated by some means. Many failure mechanisms depend on temperature, but voltage, current, and power can also be important. The appropriate acceleration factor for a specific failure mechanism is usually determined using special test structures. Examples include individual MOS transistors, ring oscillators, and special test structures that allow long chains of interconnects or metallization to be evaluated.

The situation is different for finished, packaged devices. Internal currents are usually limited to a narrow range, and the presence of overload protection structures on I/O terminals severely restricts the range of electrical conditions that can be used. Consequently, temperature is the most direct way of applying an acceleration factor on packaged parts.

4.2 High-Temperature Testing (Burn-in)

Burn-in testing has been used for many years as a means of screening marginal devices. It assumes (1) that failure mechanisms can be accelerated by raising the temperature, and (2) that the failure probability of the overall population is initially much higher, due to infant mortality.

The Arrhenius equation, shown below, is often assumed to describe the temperature dependence of the failure rate (FR):

$$FR = Ae^{-\frac{E_A}{kT}} \quad (3)$$

where A is a constant, E_A is the activation energy for the specific mechanism, k is the Boltzmann constant, and T is absolute temperature.

Time and temperature are related by the activation energy. This approach works well for cases where the activation energy is known, but its effectiveness is questionable for complex packaged devices, where failure can be caused by many different mechanisms, with different activation energies. A default assumption that is often used is that the activation energy is 0.7 eV, but activation energies for various mechanisms can range from -0.1 to 1.2 eV.

Figure 4.2-1 shows the effective increase in operating time—the acceleration factor—for three different activation energies. For the lowest activation energy, the acceleration factor at 125°C is approximately 600. Much larger acceleration factors result for higher activation energies.

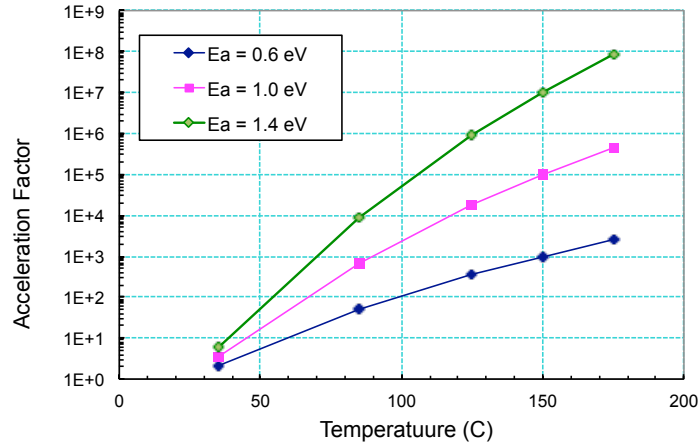


Figure 4.2-1. Acceleration factor vs. temperature for failure mechanisms with various activation energies.

A typical average activation energy for integrated circuits is often assumed to be 0.7 eV. A lower value, 0.6 eV, is recommended for more advanced processes [14]. The impact of a decrease in activation energy is to decrease the effective lifetime of burn-in at a given temperature. In order to obtain the same overall improvement, it is necessary to either increase the temperature or extend the burn-in time.

If only temperature is used to accelerate failure mechanisms, the lower activation energy means that the effective strength of infant mortality screening is reduced. For example, if 0.7 eV is valid, and the application temperature is 55°C and the burn-in temperature is +125°C, the acceleration factor is about 78. On the other hand, if the activation energy is 0.6 eV, then the acceleration factor is about 42. To achieve the same screening effectiveness, the burn-in temperature must be increased to about 140°C.

For many VLSI devices, particularly those optimized for speed, the junction temperature must be restricted to lower ranges than for previous generations and technology nodes. Therefore, raising the burn-in temperature may not only be risky, but may also increase the chance of inadvertent thermal runaway during burn-in [Ref. 1, pp. 27 and 45]. It is also possible for undetected transients during burn-in to damage good devices, increasing the chances of “walking wounded” devices being used in the flight lot. Therefore, raising the junction temperature near the absolute maximum junction temperature specified by the manufacturer for VLSI devices is not recommended.

It is far more prudent to keep the junction temperature at least 20 degrees below the absolute maximum specification and instead use voltage acceleration to cull out infant mortals. This has the significant additional advantage of removing infant mortals that are more sensitive to voltage acceleration than thermal activation. A voltage acceleration factor of between 2 and 5 is advisable for this purpose [Ref. 15, see example on pp. 9 and 10]. Using a voltage acceleration factor of 3 and a nominal electric field of 2.5 megavolts per centimeter across the gate oxide, a 5% increase in power supply voltage during burn-in is equivalent to an acceleration of 1.45. A 10% increase in power supply voltage during burn-in is equivalent to an acceleration of 2.1.

Since voltage acceleration factors are strongly technology dependent and the JEDEC document [15] is comparatively old, it should not be relied on to calculate voltage acceleration factors for the latest technologies. Recent technical literature and the VLSI manufacturer should be consulted to provide a better acceleration factor. However, a voltage acceleration of 2 (or higher) is feasible for advanced VLSI technology and offers a superior infant mortality screen.

4.3 Electrical Testing

Electrical testing can be an effective way to screen devices. Extending electrical tests to incorporate more realistic conditions that encompass actual use conditions is often done, as well as implementing special tests to determine pattern sensitivity or maximum frequency limitations.

Another approach is to develop special tests that are intended to establish that there is adequate margin in circuit functionality. For example, SRAMs can be tested at various power supply voltages, examining the number of failed bits within the array when the power supply voltage is decreased. Specific parts that have less margin can be removed, ensuring that adequate internal design margin is present, and that there are few internal defects that affect the part.

5.0 DERATING FACTORS

Derating is a method to improve reliability of parts in a space application environment by consciously reducing stresses on the device. It presumes that the relation of reliability to these stresses is known or may be established. This assumption is becoming weaker with advanced VLSI.

As mentioned previously, the Arrhenius equation is still considered valid with scaled CMOS; however, the activation energy is regarded as decreasing with more recent technology nodes. Since life testing at three temperatures (the most definitive method to establish activation energy) has proven to be prohibitively expensive in many instances, the VLSI user must assume that the activation energy of 0.5 to 0.65 eV is still valid and act accordingly (unless data is provided by the VLSI manufacturer). The traditional derating of junction temperature to 110°C is still prudent, but the 40°C margin below absolute maximum junction temperature specified by the part manufacturer may become the more critical circuit limitation. Power dissipation density is increasing in the more aggressively scaled VLSI CMOS technologies, and internal hot spots during ordinary operation are becoming more frequent. Further, variations in thermal resistance in more advanced packaging (particularly plastic packaging, which is notorious for being a poor heat conductor) is expected to exacerbate this problem.

As mentioned previously, derating voltage is considered a risky strategy for advanced VLSI parts. Typically, several voltages are generated internally in such chips and many different types of transistors of varying geometries are used to gain performance advantage. Reducing the voltage even to the lower half of the recommended operating range may result in the unintended consequence of peculiar functionality in complex devices. Such oddities may not become obvious in all circuits until the mission is launched.

A safer restriction is to derate the frequency of operation. The traditional derating of 80% is still appropriate, since the active power dissipation will thereby be also reduced to 80% uniformly within the complex VLSI device.

Generally, it is not possible to reduce currents within the VLSI other than by reducing power.

6.0 COMPARISON OF HIGH-RELIABILITY AND COTS MANUFACTURERS

Traditionally, space systems have used integrated circuits produced by manufacturers subjected to frequent surveys and review of their control systems by DLA Land and Marine, NASA, and similar organizations. ISO certification is also demanded. Manufacturers who produce only commercial devices are generally avoided, but there are exceptions, particularly large memory devices such as DRAMs that are only available commercially. This may still be a feasible strategy in the near future, but it is likely to become a greater issue. More attention needs to be given to commercial devices in the future, particularly for highly scaled devices where fundamental manufacturing is done only by commercial producers.

Due to the high cost of wafer fab (about \$2 billion and more), most wafer fabs are prioritized or monopolized for high volume commercial devices. Nevertheless, military and space integrated circuit manufacturers employ special techniques to get higher reliability wafers out of these fabs. Such techniques include selecting wafers using the results of electrical tests on test structures, and reliability tests on test structures.

EEE-INST-002 [12] is a GSFC document frequently used by space agencies, including the Jet Propulsion Laboratory (JPL), to delineate screening and qualification requirements applicable to parts used in space hardware. For traditional missions, high reliability parts controlled by Source Control Drawings are considered acceptable. For these parts, the pedigree implies and typically requires:

1. 100% screening including temperature cycling and burn-in
2. Qualification by extended temperature cycling (100 cycles versus 10 cycles for screening)
3. Life test (1000 hours instead of the 160 to 240 hours usually used for screening)
4. Various package tests such as thermal and mechanical shock
5. Internal atmosphere (water vapor) tests on samples from the flight lot

Qualification tests are still imposed on a sacrificial sample of units from the flight lot in order to provide heightened confidence in the long-term reliability of flight lot parts used for the missions.

Several issues arise in reviewing these tests when applied to advanced technology VLSI/CMOS devices. As mentioned previously, advanced devices may not allow the higher junction temperatures associated with traditional space requirements. Also advanced packaging may not be capable of the extended temperature cycling regimes demanded within EEE-INST-002. Adjustment of the burn-in (and life test) temperature may be necessary, which probably means longer test duration. However, validation of the temperature acceleration factor is necessary in such cases. In addition, voltage acceleration should seriously be considered. Burn-in and life testing at a higher voltage (in order to accelerate certain failure mechanisms), must nevertheless be restricted within the allowable power supply voltages recommended by the manufacturer. These tests may be performed, for example, at the nominal power supply voltage plus 5%.

Similarly, more temperature cycles may be needed if the non-operating thermal range is reduced. It is important, in that case, that the non-operating thermal range be compared to mission usage.

7.0 SPECIFIC RECOMMENDATIONS

7.1 Relationship of Foundries and End-Manufacturers

As discussed earlier, various options need to be considered for reliability of advanced CMOS devices in space. One of the most important factors is whether the devices are produced by space-qualified manufacturers, or by commercial foundries. Although dedicated radiation-hardened foundries are available with feature sizes as short as 150 nm, they are not available for smaller feature sizes. Radiation tolerant processes, including hardened-by-design processes, are available at the 90 nm node, but they actually use commercial foundries, not processing lines that are specifically designed to be hardened to radiation.

Another approach that has been developed is the *Trusted Foundry*, where government and national laboratory personnel develop a working relationship with the foundry that provides—through a careful process—access to proprietary information about the processing, design rules, and reliability data. The foundry agrees to make such data available for an extended period in order to allow custom circuits to be fabricated with that process for military, space, and other critical government applications. Periodic audits are made to ensure that the foundry complies with the overall requirements, and that proprietary information is properly safeguarded.

As a result, there are at least three different ways to obtain parts for space use through commercial foundries, which are the only option for highly scaled processes:

1. Parts manufactured through working agreements with mainstream producers of high-reliability parts (in this case, the arrangement between the foundry and producer depends on contracts between them)
2. Trusted foundries, as discussed above, where the government establishes the relationship between the specific foundry and the designers
3. Parts produced on modified commercial processes, where changes are made to the process for specific wafers or wafer runs that are used for high-reliability (often radiation-hardened processes)

7.2 Source Selection, Qualification, and Screening

Regardless of the specific approach used for the relationship between the end-manufacturer and foundry, three distinct steps are needed in the overall process of selecting and qualifying advanced CMOS parts for space applications.

1. Establishment of the source selection requirements for the manufacturer
2. Determination of qualification requirements for testing and evaluating the parts produced by the process, which may include monitoring process control or reliability test vehicles. These tests are generally destructive.
3. Specific screening tests on the final product that include tests directly related to the circuit application and the environmental requirements. These tests are applied to all flight parts, and must be done under carefully controlled conditions. They can never be destructive, but may result in elimination of some parts from the flight lot that do not pass the screening tests.

7.2.1 Source Selection

Source selection requirements are largely based on methods used by the manufacturer to characterize, validate, and control the various processing steps involved in manufacturing. These include dealing with statistical variations in the characteristics of individual transistors on very large-scale circuits, specifying and controlling the defect density at various steps during processing, and specific tests that are made to

Table 7.2-1. Tests and evaluations for source selection.

Mechanism	Approach Used by Vendor	Additional Steps and Processes for Space Use	Specific Screening Methods
Front-end processing	Specific failure mechanisms evaluated with test structures Statistical process evaluation	Review and track reliability data from vendor	None for this category
Back-end processing	Specific failure mechanisms evaluated with test structures (metallization, bonding and vias)	Review and track reliability data from vendor	None for this category
Packaging	Packaging yield and failure mechanisms	Review and track reliability data from vendor Construction analysis	Additional temperature cycling tests
Overall process reliability	Yield of final product (relative values may be adequate for this evaluation)	Evaluation and additional electrical testing of final devices by customer	Additional temperature cycling test Additional burn-in testing

ensure that the dominant reliability mechanisms are adequately monitored and controlled. Most of the latter tests are done on special test structures.

Table 7.2-1 lists various tests that are done during processing and assembly. Front-end processing tests are typically done using test structures. They include evaluation of hot-carrier effects, time-dependent dielectric breakdown (involving the gate and lateral insulators), and determination of the effect of drain-induced barrier lowering on short channel effects. These tests serve two purposes. Initially, they are done as part of the overall characterization of the process, and its variability, to establish design rules. Once the process is established, the tests are usually repeated periodically to ensure that the overall process remains within the expected tolerance range. Other specific tests are done, including measuring the gate oxide thickness and various sheet resistance values that also track the process. The results of these tests are often evaluated statistically, and monitored to establish statistical control boundaries.

Other tests are done to evaluate back-end processes (typically metallization and contacts). A different set of tests is established to evaluate assembly and packaging, usually including burn-in, although the conditions for burn-in at this stage may be less severe than required for space qualification.

Overall process reliability can be evaluated by tests on the final product. The customer often performs these tests after additional thermal cycling and electrical tests at various temperatures. Note, however, that the fundamental failure mechanisms due to front-end or back-end processing usually do not manifest themselves at this stage, partly because of the difficulty of determining the actual failure mode within a complex part with eight or more levels of metallization.

7.2.2 Qualification Tests

Tests at this level are usually destructive, and are done on samples from the lot used to assemble parts intended for flight use. A sample size of 22 to 45 units (based on lot-tolerance percent defective statistics) is often recommended for each test, but it may be necessary to use much smaller sample sizes for advanced parts because of their high unit cost.

Typical qualification tests are shown in Table 7.2-2. They include life tests at high and low temperature, destructive physical analyses, unbiased temperature cycling tests that are intended to evaluate weaknesses in packaging, and X-ray tests. Low-temperature life tests are a new requirement that is recommended for advanced devices because of increased concern about hot-electron degradation. Performing these tests on complete packaged parts does not provide the fundamental information of similar tests on test structures

that are usually done over a range of conditions, but it involves an extremely large number of individual transistors on the circuit, with improved statistics.

A list of various screening tests is shown in Table 7.2-3. Burn-in is nearly always performed, although the bias conditions, time, and temperature must be carefully selected to ensure that the conditions will be effective in weeding out infant mortality, and that the temperature of the die is within the limits needed to ensure that reliability is not adversely affected. This is not always straightforward for large packages because of thermal gradients. The uniformity of the die attachment and conduction paths to heat sinks can affect this, and generally require extensive analysis for large-scale devices.

Temperature cycling for 100% of the parts in the flight lot may be recommended for some complex packages. The number of cycles and the temperature range are more restricted than those used in the sample tests during qualification, which do not go into flight hardware.

Special measurements are not always used, but they can be effective in screening out marginal devices. Examples include voltage margin tests (using reduced power supply voltage to ensure that the part will operate with reduced internal logic switching voltage), and evaluation of statistical distributions of normal electrical parameters to weed out parts with atypical values, even if they are within the overall electrical specification limits. The latter approach requires “read and record” data for all parts in the flight lot.

Table 7.2-2. Qualification tests.

Test	Purpose(s)	Issues	Special Evaluations
High-temperature life test	Evaluation of temperature-activated mechanisms	—	Drift in electrical parameters Effects on TDDb, surface inversion, and electro-migration
Low-temperature life test (As appropriate)	Electrical and die reliability evaluation at low temperature	—	Effects of hot carrier degradation on overall electrical performance Drift in electrical parameters
Destructive physical analysis (DPA)	Evaluate die and construction	—	Residual gas analysis when appropriate
Extended temperature cycling	Mechanical stress on die and package after repeated deep cycling	—	Cracks at edge of large packages
X-ray	Evaluate bonding, die attach, and package assembly details	—	—

Table 7.2-3. Screening tests.

Test	Purpose(s)	Issues	Special Evaluations
Burn-in	Weed out parts affected by infant mortality	—	Weed out defective cells or bits
Temperature cycling	Weed out early package failures	Cycles must be limited to avoid impacting flight lot reliability	—
X-ray	Evaluate bonding, die attach, and package assembly details	—	—
Special electrical tests	Characterize and weed out parts with reduced operating margins	—	Verify operation under conditions for cold temperature environments

7.3 Applications in Extreme Environments

Some applications require that parts be used well beyond the normal region where they are designed and tested by the manufacturer. For example, there are many cases where parts have to be used at low temperature, such as Mars surface missions, missions to cold bodies (such as asteroids), and in applications on conventional spacecraft outside the normal electronic enclosures, where lower temperatures are encountered.

The first issue that must be dealt with is that of determining whether the part can actually function properly at temperatures well below the normal design range. Laboratory characterization tests and modeling can be used to determine this. The second issue is that of packaging. A number of mechanisms are involved, including the mismatch of thermal expansion coefficients of leads, feed-throughs, the semiconductor die, and the package. Test and qualification methods need to be developed that take these mechanisms into account.

It is important to distinguish between missions where the part is only cooled once, and cases where it must be frequently cycled between low and moderate temperatures (such as the Mars surface). Test and qualification methods are quite different for these two scenarios.

8.0 SUMMARY

This document provides recommendations for using advanced CMOS devices, with feature size below 90 nm, in space applications. The very high cost of semiconductor processing lines that are capable of producing such advanced devices makes it impossible to use dedicated foundries for the limited market in space. Thus, commercial fabrication lines have to be adapted for space use. Although one can argue that devices in space require higher reliability, this is largely offset by the increased derating factors in space. In addition, most space applications actually have lower temperature requirements than many commercial applications, including those in automobiles. There are exceptions, including surface exploration missions, as well as cases where electronics are used outside electronic enclosures, with wider temperature requirements; these applications have to be treated as special cases.

Commercial manufacturers have met the increased challenges of maintaining low defect density and high yield, despite the increased difficulty that arises with devices that have very small feature size and extremely high numbers of devices on a single chip. New approaches have been developed to deal with defects, including “hard wired” changes after wafer probing to map around defective sub-circuit regions from a complex part, and the incorporation of error correction methods within the overall functionality.

There is little that the end user can do to evaluate front-end processing mechanisms, other than working closely with manufacturers to monitor the approaches that they use to deal with them, along with statistical information about yield and the properties of test structures that are used for overall control of the manufacturing process. That is the specific approach that is recommended for these types of mechanisms.

Back-end processes are also complex, and are difficult for the end-user to evaluate. Just as for front-end processes, the most effective method is to work closely with manufacturers to understand how they monitor and control these processes, rather than attempting to deal with them at the packaged part level.

From the standpoint of packaged devices, reliability problems associated with packaging, testing, and screening are topics that can be evaluated by the end user. These include:

1. X-ray screening to evaluate die attach methods as well as issues involved with packaging
2. Burn-in testing, subject to assumptions about failure mechanisms and activation energies
3. Extending electrical tests to include read-and-record parametric tests after burn-in or other tests
4. Electrical tests using lower (or higher) power supply voltage to determine overall circuit performance margins
5. Special tests at extreme temperatures (including temperature cycling), if required by the specific application
6. Destructive physical analysis

The most difficult problem for highly scaled devices is that of dealing with new failure mechanisms, as well as with larger statistical variations in processing and device parameters. Based on recent publications, mainstream manufacturers appear to have reduced the defect density to even lower values, consistent with the requirements for very large density circuits, at least to nodes as small as 45 nm. However, it is possible that testing or field failures may occur due to the extremely complex processing, and large numbers of contacts and metallization layers. Prospective users of these technologies must carefully follow trends and examples in the literature relating to reliability, as well as investigations by NASA or the Department of Defense (DoD) related to reliability.

As devices are scaled below the 45 nm node, new challenges occur in testing and manufacturing that may not be covered by this guideline. Periodic updates are recommended to ensure that the information represents present-day technology, and that specific examples and “lessons learned” are incorporated.

9.0 REFERENCES

- [1] M. White, “Scaled CMOS Technology Reliability Users Guide,” final report for 2009 work under NASA WBS 724297k4043, available at <http://nepp.nasa.gov>.
- [2] Jet Propulsion Laboratory, Field Programmable Gate Array (FPGA) Approval Standard, Rev. 0, Process Owner: Robert Menke, Document Owner: Douglas Sheldon, Effective: Oct. 2, 2009
- [3] J. Heidecker, NEPP 2011 NAND Flash Memory Qualification Guideline, available at <http://nepp.nasa.gov>.
- [4] Y. Taur, et al., “CMOS Scaling into the Nanometer Region,” *Proc. of the IEEE*, **85**, pp. 486–504 (1997).
- [5] G. Shahidi, “Challenges of CMOS Scaling at Below 0.1 μm ,” 12th Int. Conf. on Microelectronics, pp. 5–8 (2000).
- [6] D. J. Frank, et al., “Device Scaling Limits of Si MOSFETs and Their Application Dependencies,” *Proc. of the IEEE*, **89**, pp. 259–288 (1999).
- [7] D. Duarte, et al., “Impact of Scaling on the Effectiveness of Dynamic Power Reduction,” *Proc. of the 2002 International Conf. on Computer Design: VLSI in Computers and Processors* (2002).
- [8] K. J. Kuhn, et al., “Process Technology Variations,” *IEEE Trans. Elect. Dev.*, **58**(8), pp. 2197–2208 (2011).
- [9] F. Chen, et al., “Critical Ultra Low-k TDDB Reliability Issues for Advanced CMOS Technologies,” *IEEE Reliability Physics Symposium*, pp. 464–475 (2009).
- [10] P. Lall, et al., “Leading Indicators of Failure for Prognostication of Leaded and Lead-Free Electronics in Harsh Environments,” *IEEE Trans. Comp. and Packaging Technology*, **32**(1), pp. 135–144 (2009).
- [11] S. Manson, *Thermal Stress and Low Cycle Fatigue*, McGraw-Hill: New York, NY, 1966.
- [12] EEE-INST-002: Instructions for EEE Parts Selection, Screening, Qualification, and Derating NASA/TP—2003–212242.
- [13] Jet Propulsion Laboratory, JPL Standard: Part Engineering Technical Standard DocID 78157.
- [14] J. Srinivasan, et al., “The Impact of Technology Scaling on Lifetime Reliability,” Int. Conf. on Dependable Systems and Networks, June 2004.
- [15] JEDEC Document JEP122E, “Failure Mechanisms and Models for Semiconductor Devices,” March 2009.